DOCUMENT RESUME

ED 087 414 IR 000 158

AUTHOR	Joos, L. W.
TITLE	Computers Analysis of Reading Difficulty.
PUB DATE	Apr 73
NOTE	4p.; Paper presented at the Association for Educational Data Systems Annual Convention (New Orleans, Lousiana, April 16 through 19, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Computer Programs; *Content Analysis; Program Descriptions; Reading; *Reading Difficulty; Reading Level; Reading Material Selection; Reading Programs
IDENTIFIERS	AEDS; Association for Educational Data Systems; *Dale Chall Formula

ABSTRACT

A computer program has been designed to analyze the reading difficulty of English text. It is, essentially, an automation of the widely used Dale-Chall formula for the estimation of reading difficulty. Text samples of up to 5000 words are input, with only minor punctuation restrictions being applied. Output from the program consists of sample analyses, a summary of sample statistics and a tally of all different words in the sample text. Basic data used by the Dale-Chall formula are counts of sentences and of common and uncommon words. The analyses include counts of words and sentences, a frequency listing of words used, the Dale-Chall value and tabled reading length, and several statistics such as sentence length and percent of uncommon words. The program can be used for reading and library projects, for instructional materials analysis and selection, and is of interest in any field where published text is expected to be read. (PB)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ED 087414

COMPUTER ANALYSIS OF READING DIFFICULTY

L. W. Joos

Oakland Schools (Mich.)

INTRODUCTION

This paper describes the theoretical basis for and the salient characteristics of a computer program for analyzing the reading difficulty of English text. Basically, the system consists of the automation of a procedure long used in education, which makes use of certain research conducted by Dale and Chall (1). Input, output, and library data sets are described. Functions performed by the computer are briefly described.

Theory

Readability is a term used to describe the degree of difficulty of text material for the person who is to read it. Since difficulty is a concept that relates to people, reading difficulty must be expressed in terms of the abilities of various kinds of people who will attempt to read text.

One commonly accepted continuum for reading ability is that which corresponds to school grades. Thus we say that a student reads comfortably at the 6th grade level, but has some difficulty with 7th grade text. Conversely, 6th grade text is considered to be that text which can be read by the average 6th grade pupil, but probably not by a 5th grader of average reading ability. The argument is very nearly circular, in that we define reading difficulty in terms of ability, and ability in terms of difficulty. Using this circular argument, it is possible to entirely avoid questions which have to do with the parameters of difficulty, or for that matter the parameters of reading ability. What things make text hard to read (or easy) obviously are related to what reading skills people have; but the reading skills that people have are necessarily developed by exercise in reading.

A useful assumption is that reading skills are very much like other learned skills in that the population of readers may be thought of as possessing ability which gradually increases with maturation. Briefly, reading is a developmental skill, which can be laid on the grade level scale with reasonable success. Equally well, reading difficulty of text may use the same scale.

If now we consider the question of determining the readability of a given text, we usefully define it as corresponding to the grade level of

---

DR. L. W. JOOS is Director of Systematic Studies, Oakland Schools, Pontiac, Michigan. He is a graduate of the Universities of Minnesota and Wisconsin, and has been conducting research in educational applications of computer technology since 1961. He is the programmer and designer of several such applications, including PACER, ASK/1, READABILITY SCORER (copyright, Oakland Schools).

IR000 158

the reader who could not read it easily if it were more difficult. We may then wish to determine whether a given sample of readers can comfortably read a certain text by a practical trial -- that is, allow the readers to attempt to read the text. If, under such a trial, we determine by inquiry or testing that the average readers in the sample had been successful in reading the text in question, the readability of the text is now defined by the characteristics of the successful readers.

Unfortunately, practical trials in which samples of readers attempt to read samples of text are expensive in time and money, even though they may well be the ultimate test of readability. Alternatively, we turn to an investigation of the text material itself, to see whether certain characteristics of English text seem correlated with reading difficulty. If the text characteristics which we find to be sufficiently related to reading difficulty are themselves simple and easily assessed, the estimation of readability becomes a matter of mechanically scanning text. That is essentially what has been done by the Joos-Butz Readability Scorer.

Research Basis

Basic research in the field of measurement of readability has been very extensive, and goes back many years. The first true readability formula was published 50 years ago by Lively and Pressey (2). Klare (3) lists 482 entries in his bibliography, 79 of which are concerned with formulas. The formula developed by Dale and Chall stems from their attempt to improve upon previously developed formulas.

The Dale-Chall formula appeared in 1948, and has become perhaps the most widely used. Since it involves only two parameters, both of which can be obtained from text by computer scanning, the Dale-Chall formula is well suited to computer processing. It is also widely accepted and widely known.

The formula is actually a linear regression obtained from a research study, and yields an unbiased estimate of a value called X_{c50}. This is defined as the reading grade level of a student who could answer one-half of the test questions on a passage from which the text sample was taken. The computed value is looked up in a table which applies certain corrections and which also maps the value onto the grade equivalent scale for easy interpretation.

Joos-Butz Readability Scorer

The work of Dale and Chall had reduced the estimation of readability to a task which could be programmed for a computer. The input to the computer program consists of one or more samples of English text, usually key-punched into cards. The text is punched in essentially the original format including punctuation.

Each text sample may be as short as a single sentence, or as long as 5,000 words. Most samples run in the 500 - 1000 word range. Two limitations are essential as to the use of punctuation: the character '.' may be

used only as a sentence marker (not in abbreviations or decimal fractions, etc.); asterisks are used only to mark proper names. Two non-English words are used to mark the beginning and end of samples: BEGSAMP, and ENDSAMP. However, these words are markers only when found in the first positions of an input record.

Output from the program consists of sample analyses, a summary of all samples statistics, and a tally of all different words in the sample.

The sample analysis includes a copy of the input, a running count of words and sentences, an alpha-sort listing of the words used with the frequency of each, the Dale-Chall value and tabled reading level, and several statistics including average sentence length, average not-common words per sentence, and percent different words.

The basic data used by the formula consists of a sentence count, a word count, and a count of words not found in a basic list of 8,000 so-called common words. The basic list when expanded for computer use totals about 13,000 words, since all forms of a word must be included. This list is stored in memory as a file of 13,000 records each 21 characters in length, thus occupying over 270,000 bytes of storage.

By sorting the sample into alphabetic order, and by looking up each different word only once, the determination for each word as to its existence in the list takes a minimum of time. The common word file is stored in keyed indexed environment and is maintained as a catalogued data set in the system.

Each input sample is stored in core in an array of 5,000 (maximum) elements. The program has an in-line sorter of this array, since this seemed more convenient than using the operating system's sort utility. As each sample is processed, the sample array is written to a sequential file containing only one record for each different word, but with the frequency and sample number attached.

At the conclusion of processing all samples in the job, the accumulated file of words is sorted by the sort utility and returned to a program which prints the total vocabulary for the whole run.

There are two critical programming problems in this system. The first, already discussed, is that of maintaining the common word list in fast access mode. Without the existence of keyed indexed files in random storage hardware, this would be impossible.

The other critical problem is the logic involved in scanning the sample text. Words are identified by word-markers, which include blanks, commas, colons, semi-colons, and sentence markers. Sentence markers are periods, question marks, and exclamation points. Provision is also made for recognizing comments inserted in the sample. The language used is PL/1 which is ideally suited to problems of this kind. Currently there is a project under way to write the program in assembler language for use in DOS systems.

Applications of the Readability Scorer have been many. Some jobs have been run by our teleprocessing users who utilize 2741 typewriter terminals for text input. The largest application has been its use in Reading Power Project, which has analyzed whole libraries of text books for use in vocational schools. Printer format output from the above project is stored in disk packs and includes as many as 50 books on a pack.

While the superiority of automating the Dale-Chall formula is obviously great in comparison to the people-powered table look-up that they used, the research in this area is not finished. We anticipate the possibility of by-passing the key-punching in various ways. We further anticipate that other assessible characteristics of text may be practically utilized by computer system. In the meantime, the current system is wonderfully useful and efficient. We expect an expansion in many fields, in education and wherever published text is expected to be read.

SUMMARY

This paper has briefly described the characteristics of a computer program for use in analyzing English text material by use of the Dale-Chall formula. The system has been proved in use and will be of interest to computer installations in colleges and universities throughout the English speaking world. The language used in the current version is PL/1, and the hardware used is IBM 360/50 with IBM 2314 disc storage. The PL/1 program is not compatible with IBM DOS.

BIBLIOGRAPHY

1. Dale, E., and Chall, J. S. "A Formula for Predicting Readability", Educational Research Bulletin, 27:11-20, January 21, 1948.

2. Lively, B. A., and Pressey, S. L. "A Method for Measuring the 'Vocabulary Burden' of Textbooks", Educational Administration and Supervision, 9:389-98, October, 1923.

3. Klare, George R. "The Measurement of Readability", Iowa State University Press, 1963.